International Journal of Machine Learning and Cybernetics (2024) 15:2737–2747
https://doi.org/10.1007/s13042-023-02060-9

ORIGINAL ARTICLE

Fuzzy information gain ratio-based multi-label feature selection with label correlation

Ying Yu[1,2] · Meiyue Lv[2] · Jin Qian[2] · Jingqin Lv[2] · Duoqian Miao[3]

Received: 4 October 2023 / Accepted: 1 December 2023 / Published online: 21 January 2024

Abstract
Multi-label feature selection aims to mitigate the curse of dimensionality in multi-label data by selecting a smaller subset of features from the original set for classification. Existing multi-label feature selection algorithms frequently neglect the inherent uncertainty in multi-label data and fail to adequately consider the relationships between features and labels when assessing the importance of features. In response to this challenge, a Fuzzy Information Gain Ratio-based multi-label feature selection considering Label Correlation (FIGR_LC) algorithm is proposed. FIGR_LC evaluates feature importance by combining the relationship between features and individual labels, as well as the correlation between features and label sets. Subsequently, a feature ranking is established based on these feature weights. Experimental results substantiate the effectiveness of FIGR_LC, showcasing its superiority over several established feature selection methods.

Keywords Multi-label feature selection · Fuzzy information gain ratio · Fuzzy rough sets

1 Introduction

In conventional supervised learning, a sample is usually associated with only one category label, which represents one specific semantic meaning [1]. However, real-world scenarios often involve objects with multiple semantic information, potentially belonging to multiple categories simultaneously [2]. For instance, a text can be both political and economic; An image might include various semantic items, such as meadow and flower; In the gene function prediction, a gene might be involved in both transcription and metabolism at the same time. Clearly, these multi-sense objects cannot be accurately described by a single category label. This challenge poses a significant problem for traditional supervised learning, which emphasizes singular and distinct semantic meanings. To address this challenge, the multi-label learning framework has been introduced [3], in which each object is associated with a set of class labels that represents multiple semantic information. It could leverage multi-label training datasets to predict the class labels for unlabeled multi-label samples and finds widespread applications in various domains. Specifically, it is extensively utilized in text classification [4, 5], automatic image and video labeling [6–8], as well as scene classification [9].

The widespread availability of high-dimensional multi-label data potentially leads to the curse of dimensionality and compromising the accuracy of multi-label learning. To address this, numerous multi-label dimensionality reduction methods have been proposed. These methods generally fall into two categories: feature extraction and feature selection. The process of transforming the feature space from its higher original dimensionality into lower dimensionality using mapping or transformation is referred to as feature extraction. For example, Multi-label Informed Latent

✉ Ying Yu
yuyingjx@163.com

Meiyue Lv
meiyuelv@163.com

Jin Qian
qjqjlqyf@163.com

Jingqin Lv
jingqinlv@ecjtu.edu.cn

Duoqian Miao
dqmiao@tongji.edu.cn

[1] State Key Laboratory of Performance Monitoring and Protecting of Rail Transit Infrastructure, East China Jiaotong University, Nanchang 330013, Jiangxi, China

[2] College of Software, East China Jiaotong University, Nanchang 330013, Jiangxi, China

[3] Department of Computer Science and Technology, Tongji University, Shanghai 201804, China

Semantic Indexing (MLSI) [10] is a notable feature extraction algorithm. Although feature extraction can reduce the dimensionality of multi-label feature space, it also destroy the structure of the original feature space. The newly generated feature space loses its original semantics, resulting in a classification model that cannot be fully interpreted. Conversely, feature selection focuses on selecting a subset of features from the original set based on specific evaluation criteria. This approach not only reduces the dimensionality of the original feature space by eliminating unnecessary or redundant attributes but also preserves the original semantics. As a result, feature selection has garnered more attention in research and applications, providing a balance between dimensionality reduction and retaining the interpretability and meaning of the data.

The key to feature selection is to figure out the importance of each feature and then select features with higher importance, which can generally be measured through correlation. Unlike traditional single-label learning, each multi-label object may have multiple semantic labels simultaneously, so multi-label feature selection approaches should consider not only feature-to-label and feature-to-feature correlations, but also feature-to-label set correlations, as well as correlations between labels [11, 12]. Currently, some existing multi-label feature selection algorithms can describe the correlation well based on some valid evaluation metrics, such as information entropy [13–15], dependency [16], or classification interval [17, 18]. For example, three multi-label feature selection algorithms, including NFNMI-opt, NFNMIneu, and NFNMIpes, were proposed based on pessimistic, neutral, and optimistic neighborhood information entropy by Lin et al. [13]. Li et al. [14] proposed a information gain-based multi-label feature selection algorithm, IGML, which uses the maximum information gain to measure the relation between features and class labels. Reyes et al. [18] presented three extensions of the popular feature estimation algorithm, ReliefF, for multi-label feature selection algorithms, namely ReliefF-ML, PPT-ReliefF, and RReliefF-ML. However, most of the existing algorithms mainly consider feature-to-label correlations or feature-to-feature correlations.

Similar to the traditional single-label feature selection, multi-label feature selection also faces the challenge of uncertainty, including randomness, ambiguity and inconsistency. As a valuable technique for analyzing data uncertainty [19], rough sets theory [20] can describe the dependencies within the data under the condition of limited information granularity and is therefore regarded as an effective tool for dimensionality reduction [21–24]. Originally, rough sets was ill-suited for handling continuous data, leading to their extension to overcome this limitation. One of the primary extension model for rough sets is fuzzy rough sets, which replaces equivalence relations with similarity relations in classical rough sets to calculate indistinguishability. Several evaluation metrics based on fuzzy rough sets were introduced, including fuzzy dependency functions [25–28] and fuzzy information entropy [29, 30]. These metrics paved the way for the creation of a fuzzy discernibility matrix for single-label feature selection [31, 32]. Subsequently, it was extended to multi-label feature selection, resulting in numerous algorithms based on fuzzy rough sets [33–36]. Zhang et al. [33] proposed an multi-label feature selection algorithm by amalgamating dependency functions with BR as a fuzzy rough set-based feature assessment. Lin et al. [34] introduced two algorithms for multi-label feature selection, leveraging the stream features of multi-label data and fuzzy mutual information, separately. Li et al. [36] executed multi-label feature selection by combining fuzzy rough set and multi-kernel learning, introducing a model utilizing kernelized fuzzy rough sets (RMFRS). However, it's important to note that these approaches only partially assess the correlations between features, leaving room for further exploration and refinement.

To address the aforementioned problems, a novel fuzzy information gain ratio-based multi-label feature selection algorithm considering label correlation (FIGR_LC) is proposed. FIGR_LC considers the correlation between features and each label, as well as the correlation between features and label sets that incorporate inter-label correlation. It utilizes the fuzzy information gain ratio to measure the correlation and defines label weights accordingly. By integrating correlation measurement with label weights, the proposed algorithm established the significance of features. Subsequently, features are ranked based on their weights in descending order, creating the feature ranking. Extensive experiments confirm the efficiency of FIGR_LC. The contributions of this research can be summarized as follows:

1. A fuzzy information gain ratio-based multi-label feature selection algorithm with label correlation is proposed, which employs fuzzy rough set information theory to handle the ambiguity and uncertainty inherent in multi-label data.
2. Different from the existing multi-label feature selection algorithms which partially consider the relationship between features and labels, the proposed algorithm considers the relationship between features and labels in terms of individual label-feature associations as well as the relevance between the feature and label sets, which also takes into account the label relevance.
3. The proposed algorithm employs fuzzy information gain ratios to quantify relevance and establish label weights. By integrating both correlation and label weights, the algorithm determines the importance of features. This importance assessment guides the derivation of a feature ranking.

The rest of this paper is organized as follows: Sect. 2 reviews the fundamental concepts of multi-label learning and fuzzy rough sets. Section 3 explores various aspects of FIGR_LC. Section 4 presents the detailed methodology of FIGR_LC. Section 5 describes experimental findings. Finally, Sect. 6 provides the concluding remarks.

2 Preliminaries

This chapter introduces the fundamental notions used in this paper.

2.1 Multi-label learning

Given a sample space $\mathfrak{X} = \mathbb{R}^{\mathfrak{q}}$, a label space with ℓ labels. Each sample $\mathfrak{x}_i \in \mathfrak{X}$ denotes a q-dimensional feature vector, $\mathfrak{Y}_i = \{\mathfrak{y}_{i1}, \mathfrak{y}_{i2}, \ldots, \mathfrak{y}_{i\ell}\}$ represents the label vector related to sample $\mathfrak{x}_i$. For each label $\mathfrak{y}_j$, if $\mathfrak{x}_i$ is connected to $\mathfrak{y}_j$, then $\mathfrak{y}_{ij} = 1$; otherwise $\mathfrak{y}_{ij} = 0$.

2.2 Fuzzy rough sets

In traditional rough sets, equivalence relations are calculated from symbolic data, while in fuzzy rough sets, numeric type features can produce fuzzy equivalence relations. Let $\mathbb{U}$ be a finite nonempty universe and $\mathfrak{R}$ be a fuzzy equivalence relation on $\mathbb{U}$. For $\forall m, o, n \in \mathbb{U}$, $\mathfrak{R}$ meets the properties below:

1. Reflexivity: $\forall m \in \mathbb{U}, \mathfrak{R}(m, m) = 1$;
2. Symmetry: $\forall m, o \in \mathbb{U}, \mathfrak{R}(m, o) = \mathfrak{R}(o, m)$;
3. Transitivity:$\mathfrak{R}(m, n) \geq \min_o\{\mathfrak{R}(m, o), \mathfrak{R}(o, n)\}$.

Given a object set $\mathbb{U}$, the feature set is $\mathbb{F}$, $\gamma \in \mathbb{F}$. The fuzzy relationship matrix $M(\gamma)$ based on feature γ is denoted as:

$$M(\gamma) = \begin{pmatrix} \mathfrak{r}_{11} & \mathfrak{r}_{12} & \cdots & \mathfrak{r}_{1z} \\ \mathfrak{r}_{21} & \mathfrak{r}_{22} & \cdots & \mathfrak{r}_{2z} \\ \vdots & \vdots & \ddots & \vdots \\ \mathfrak{r}_{z1} & \mathfrak{r}_{z2} & \cdots & \mathfrak{r}_{zz} \end{pmatrix} \tag{1}$$

where the relationship between $\mathfrak{x}_m$ and $\mathfrak{x}_n$ is $\mathfrak{r}_{mn} \in [0, 1]$. In the traditional rough sets, when $\mathfrak{x}_m$ equals $\mathfrak{x}_n$, $\mathfrak{r}_{mn} = 1$, otherwise $\mathfrak{r}_{mn} = 0$.

Definition 1 [37] Let $\mathbb{U}$ denotes sample set, $\mathfrak{R}$ denotes fuzzy equivalence relation. fuzzy partition of $\mathbb{U}$ generated by the $\mathfrak{R}$ is as follows:

$$\mathbb{U}/\mathfrak{R} = \left\{\left[\mathfrak{x}_m\right]_{\mathfrak{R}}\right\}_{m=1}^{z}, \tag{2}$$

where $\left[\mathfrak{x}_m\right]_{\mathfrak{R}} = \frac{\mathfrak{r}_{m1}}{\mathfrak{x}_1} + \frac{\mathfrak{r}_{m2}}{\mathfrak{x}_2} + \cdots + \frac{\mathfrak{r}_{mz}}{\mathfrak{x}_z}$ denotes the fuzzy equivalence class. "+" denotes "union" and "−" denotes a separator. The partition $\mathbb{U}/\mathfrak{R}$ is produced by the $\mathfrak{R}$. $\left[\mathfrak{x}_m\right]_{\mathfrak{R}}$ represents a fuzzy set, as a result of fuzzy equivalence relation.

Definition 2 [37] Cardinalities of $\left[\mathfrak{x}_m\right]_{\mathfrak{R}}$ are as follows:

$$\left|\left[\mathfrak{x}_m\right]_{\mathfrak{R}}\right| = \sum_{n=1}^{z} \mathfrak{r}_{mn}. \tag{3}$$

2.3 Fuzzy rough sets information measurement

Definition 3 [29] Given an object space $< \mathbb{U}, \mathfrak{R} >$, the fuzzy information entropy of feature γ is as follows:

$$FH(\gamma) = -\frac{1}{z}\sum_{m=1}^{z} \log\frac{\left|\left[\mathfrak{x}_m\right]_{\mathfrak{R}}\right|}{z}. \tag{4}$$

Definition 4 [29] Given a information system $\varsigma =< \mathbb{U}, \mathbb{F}, \nu, \varphi >$, $\mathbb{F}$ is attribute set, ν is the attribute value range, the mapping $\varphi = \mathbb{U} \times \mathbb{F} \rightarrow \nu$. The equivalence classes $\left[\mathfrak{x}_m\right]_X$ and $\left[\mathfrak{x}_m\right]_Y$ that contain $\mathfrak{x}_m$ are produced by feature subsets X and Y. The following is fuzzy joint entropy of X and Y:

$$FH(XY) = -\frac{1}{z}\sum_{m=1}^{z} \log\frac{\left|\left[\mathfrak{x}_m\right]_X \cap \left[\mathfrak{x}_m\right]_Y\right|}{z}, \tag{5}$$

where $\left[\mathfrak{x}_m\right]_X \cap \left[\mathfrak{x}_m\right]_Y = min\{\left[\mathfrak{x}_m\right]_X, \left[\mathfrak{x}_m\right]_Y\}$.

Definition 5 [29] Given a fuzzy rough decision table $FRDT =< \mathbb{U}, \mathbb{F}, \nu, \varphi >$, $\mathfrak{I}$ is conditional attribute set, $\mathfrak{T}$ is decision attribute, $\mathbb{F} = \mathfrak{I} \cup \mathfrak{T}$, $X \subseteq \mathfrak{I}$. The equivalence classes $\left[\mathfrak{x}_m\right]_X$ and $\left[\mathfrak{x}_m\right]_{\mathfrak{T}}$ that contain $\mathfrak{x}_m$ are produced by X and $\mathfrak{T}$. The fuzzy conditional entropy from $\mathfrak{T}$ to X is as follows:

$$FH(\mathfrak{T} \mid X) = -\frac{1}{z}\sum_{m=1}^{z} \log\frac{\left|\left[\mathfrak{x}_m\right]_X \cap \left[\mathfrak{x}_m\right]_{\mathfrak{T}}\right|}{\left|\left[\mathfrak{x}_m\right]_X\right|}. \tag{6}$$

Theorem 1 [29] $FH(\mathfrak{T} \mid \mathfrak{X}) = FH(\mathfrak{X}\mathfrak{T}) - FH(\mathfrak{X})$.

Definition 6 [29] Given a fuzzy rough decision table $FRDT =< \mathbb{U}, \mathbb{F}, \nu, \varphi >$, $\mathfrak{I}$ is conditional attribute set, $\mathfrak{T}$ is decision attribute, $\mathbb{F} = \mathfrak{I} \cup \mathfrak{T}$, $X \subseteq \mathfrak{I}$. The equivalence classes $\left[\mathfrak{x}_m\right]_X$ and $\left[\mathfrak{x}_m\right]_{\mathfrak{T}}$ that contain $\mathfrak{x}_m$ are produced by X and $\mathfrak{T}$. The fuzzy mutual information of X and $\mathfrak{T}$ is as follows:

$$I(X;\mathfrak{T}) = FH(\mathfrak{T}) - FH(\mathfrak{T} \mid X). \tag{7}$$

3 Fuzzy information gain ratio and label relevance

3.1 Correlation measurement method

Information gain signifies the degree of uncertainty reduction in the random variable α once the random variable β is determined. As a metric for assessing the significance of features, information gain gauges the information a feature provides to the system. The importance of a feature is directly proportional to the information it contributes. The definition can be stated as follows:

$$IG(\alpha, \beta) = FH(\alpha) - FH(\alpha \mid \beta). \tag{8}$$

Feature selection relying on information gain often favors features with a wide range of values, sometimes resulting in the selected features meaningless. The introduction of information gain ratio mitigates this bias by factoring in the ratio of information gain to the information entropy of the feature. Hence, considering the information gain ratio for feature selection, as outlined in Definition 7, can be a more balanced and precise approach.

Definition 7 Given a sample set $\mathbb{U}$, a feature set $\mathbb{F}$, α and β are two features or feature sets describing the sample, the fuzzy information gain ratio is defined as follows:

$$IGR(\alpha, \beta) = \frac{IG(\alpha, \beta)}{FH(\beta)} = \frac{FH(\alpha) - FH(\alpha \mid \beta)}{FH(\beta)}. \tag{9}$$

In Definition 7, IGR denotes the relation between α and β. The stronger relation between features α and β, the higher value of $IGR(\alpha, \beta)$.

3.2 Correlation measures for feature and label

Each multi-label object is associated with a set of category labels, representing various semantic. In this context, not only the features exhibit correlations among themselves, but also the labels also demonstrate interconnections. These intricate relationships significantly influence the calculation of feature importance, as highlighted in previous research [34]. Solely examining the connection between features and label sets would overlook crucial information, including the relevance between individual labels and the associations between features and each label [38]. Consequently, FIGR_LC integrates the relation between labels and features with that between features and label sets. This comprehensive perspective also takes into consideration label relevance, allowing for a holistic analysis of the relationships between features and label sets.

In Definition 7, the term IGR denotes the correlation between two features, forming the basis for defining the relationship between features and labels.

Definition 8 Given a multi-label table $\mathfrak{T} =< \mathbb{U}, \mathbb{F}, \mathfrak{L} >$, $\mathbb{U}$ represents sample set, $\mathbb{F}$ denotes the feature set, $\mathfrak{L}$ denotes label set, $f_m \in \mathbb{F}$, $\mathfrak{L}_n \in \mathfrak{L}$. The correlation between f_m and $\mathfrak{L}_n$ is as follows:

$$IGR(\mathfrak{L}_n, f_m) = \frac{FH(\mathfrak{L}_n) - FH(\mathfrak{L}_n \mid f_m)}{FH(f_m)}. \tag{10}$$

In Definition 8, the larger value of $IGR(\mathfrak{L}_n, f_m)$, the greater correlation between label $\mathfrak{L}_n$ and feature f_m.

3.3 Label relevance

The weight assigned to a label signifies its significance and its interconnection with other labels. A higher label weight indicates a greater likelihood of the label's importance. Information gain is employed to gauge the relevance between two labels and ascertain the significance of these labels in relation to each other.

Definition 9 Given a multi-label table $\mathfrak{T} =< \mathbb{U}, \mathbb{F}, \mathfrak{L} >$, $\mathbb{U}$ represents sample set, $\mathbb{F}$ denotes the feature set, and $\mathfrak{L}$ denotes label set containing ℓ labels. The weight of $\mathfrak{L}_m$ is as follows, where $\mathfrak{L}_m \in \mathfrak{L}$:

$$W(\mathfrak{L}_m) = \frac{\sum_{n=1}^{\ell} IG(\mathfrak{L}_n, \mathfrak{L}_m)}{\sum_{m=1}^{\ell} \sum_{n=1}^{\ell} IG(\mathfrak{L}_n, \mathfrak{L}_m)}, \tag{11}$$

where $IG(\mathfrak{L}_n, \mathfrak{L}_m)$ denotes the information gain between $\mathfrak{L}_m$ and $\mathfrak{L}_n$. $W(\mathfrak{L}_m)$ represents the importance of label $\mathfrak{L}_m$ and is equal to the ratio of the sum of information gain of label $\mathfrak{L}_m$ and other labels to the information gain among all labels. Obviously, it can be seen that $0 < W(\mathfrak{L}_m) < 1$ and $\sum_{m=1}^{\ell} W(\mathfrak{L}_m) = 1$.

3.4 Correlation measures for feature and label sets

In order to calculate the relevance between features and label sets, which incorporates label significance, we construct a relationship matrix concerning $\mathfrak{L}$ that considers label correlations. The conventional method of computing the label relationship matrix mandates absolute identity in all labels between two samples for the value to be 1; otherwise, it defaults to 0. However, in the context of multi-label data, this traditional approach is excessively stringent. It often yields a similarity score of 0 between two samples, neglecting label relevance and failing to effectively capture the label set's characteristics.

Consequently, we propose a novel method for calculating label set relationships, integrating label weights, as delineated in Definition 10.

Definition 10 Given a multi-label table $\mathfrak{T} =< \mathbb{U}, \mathbb{F}, \mathfrak{L} >$, $\mathbb{U}$ represents sample set, $\mathbb{F}$ denotes feature set, $\mathfrak{L}$ denotes label set containing ℓ labels. $W(\mathfrak{L}_k)$ denotes the weight of label $\mathfrak{L}_k$. Then the similarity of $\mathfrak{x}_m$ and $\mathfrak{x}_n$ with respect to $\mathfrak{L}$ is defined as follow:

$$s_{mn}^{\mathfrak{L}} = \sum_{k=1}^{\ell} W(\mathfrak{L}_k)(\mathfrak{y}_{mk} = \mathfrak{y}_{nk}), \tag{12}$$

if $\mathfrak{y}_{mk} = \mathfrak{y}_{nk}$, it returns 1, otherwise returns 0. The weight information is incorporated into the similarity relation of label sets, which also means that the correlation between the labels is considered.

Since label set consists of symbolic data with a value of 0 or 1 in the relationship matrix, we convert the similarity to 0 or 1 according to Definition 11.

Definition 11 Given a multi-label table $\mathfrak{T} =< \mathbb{U}, \mathbb{F}, \mathfrak{L} >$, $\mathbb{U}$ represents sample set, $\mathbb{F}$ denotes the feature set, and $\mathfrak{L}$ denotes label set containing ℓ labels. $s_{mn}^{\mathfrak{L}}$ denotes the similarity of samples $\mathfrak{x}_m$ and $\mathfrak{x}_n$. If $s_{mn}^{\mathfrak{L}} >= \gamma$, the similarity of samples $\mathfrak{x}_m$ and $\mathfrak{x}_n$ is equal to 1. It is defined as follows:

$$\mathfrak{r}_{mn}^{\mathfrak{L}} = \begin{cases} 1, & s_{mn}^{\mathfrak{L}} >= \gamma \\ 0, & \text{otherwise} \end{cases}, \tag{13}$$

where the default value of threshold γ is 0.8.

Calculate the relationship matrix of label sets $M(\mathbb{R}_{\mathfrak{L}}) = \left[\mathfrak{r}_{mn}^{\mathfrak{L}}\right]_{z\times z}$ by Eq. (13). The information entropy of the label set is calculated based on Eq. (4) and Definition 7.

4 The proposed method

According to the preceding analysis, label correlation could offer valuable additional information to improve the efficiency of multi-label learning. Therefore, it is imperative to consider the correlation between labels in the process of multi-label feature selection.

Definition 12 Given a multi-label table $\mathfrak{T} =< \mathbb{U}, \mathbb{F}, \mathfrak{L} >$, $\mathbb{U}$ represents sample set, $\mathbb{F}$ denotes the feature set, $f_m \in \mathbb{F}$, $\mathfrak{L}$ denotes label set containing ℓ labels. The weight of feature f_m is defined as follows:

$$\begin{aligned} IGRS(f_m) &= \sum_{n=1}^{\ell} IGR(\mathfrak{L}_n, f_m) + IGR(\mathfrak{L}, f_m) \\ &= \sum_{n=1}^{\ell} \frac{FH(\mathfrak{L}_n) - FH(\mathfrak{L}_n \mid f_m)}{FH(f_m)} \\ &\quad + \frac{FH(\mathfrak{L}) - FH(\mathfrak{L} \mid f_m)}{FH(f_m)}. \end{aligned} \tag{14}$$

Based on Definition 12, a fuzzy information gain ratio-based multi-label feature selection algorithm with label correlation (FIGR_LC) is proposed. Algorithm 1 provides comprehensive instructions.

Algorithm 1 Fuzzy information gain ratio-based multi-label feature selection with label correlation (FIGR_LC)

Input: A multi-label decision table $\mathfrak{T} =< \mathbb{U}, \mathbb{F}, \mathfrak{L} >$, where $\mathbb{U}$ denotes a nonempty finite set of samples, $\mathbb{F} = \{f_1, f_2, \cdots, f_q\}$ denotes the feature set, and $\mathfrak{L}$ denotes label set containing ℓ labels

Output: The feature ranking $rank$

```
//Compute the information gain between labels
for m = 1 : ℓ do
    for n = 1 : ℓ do
        IG(𝔏_n, 𝔏_m) = FH(𝔏_n) − FH(𝔏_n | 𝔏_m)
    end for
end for
//Compute the weight of each label
for m = 1 : ℓ do
    W(𝔏_m) = (Σ_{n=1}^{ℓ} IG(𝔏_n, 𝔏_m)) / (Σ_{m=1}^{ℓ} Σ_{n=1}^{ℓ} IG(𝔏_n, 𝔏_m))
end for
//Compute the weight of each feature
for m = 1 : q do
    //Compute the correlation between f_m and each label
    for n = 1 : ℓ do
        IGR(𝔏_n, f_m) = (FH(𝔏_n) − FH(𝔏_n | f_m)) / FH(f_m)
        IGR_SUM = IGR_SUM + IGR(𝔏_n, f_m)
    end for
    //Compute the correlation between f_m and 𝔏 with label correlation
    IGRS(f_m) = IGR_SUM + (FH(𝔏) − FH(𝔏 | f_m)) / FH(f_m)
end for
Sort features in descending order of IGRS to eatablish the feature ranking rank
return rank
```

The weight of feature f_m is considered from two different perspectives. Firstly the correlation between feature f_m and each label is calculated. Then, the correlation between the feature f_m and the label set $\mathfrak{L}$ is calculated, where the correlation between the labels is considered. Finally, two results are combined as feature weight to evaluate the importance of the feature f_m.

Let $|F|$ represent the number of features, $|L|$ is the number of labels, and $|U|$ denotes the number of samples. Algorithm 1 mainly includes three steps. First, the weights of labels are calculated with $O\left(|L|^3\right)$ time complexity. Then the feature weights $IGRS(f_m)$ is calculated, with $O(|F||L|)$ time complexity of the correlation. Finally, feature ranking is performed, and its time complexity is $O(|F||F|)$. Overall temporal complexity is $O\left(|L|^3 + |F||L| + |F||F|\right)$.

5 Experiments study

In this section, the proposed algorithm is compared with five popular multi-label feature selection algorithms in terms of classification performance.

5.1 Multi-label datasets

In the experiments, five baseline multi-label datasets from a variety of fields are used to evaluate the effectiveness of the proposed algorithm. All of these used datasets are available from the *Mulan* Library. Table 1 summarises the characteristics of these datasets.

5.2 Evaluation metrics

Five evaluation metrics [33] are used to assess the algorithm performance and their particulars are as follows:

Let $Z = \left\{\left(x_i, Y_i\right)\right\}_{i=1}^{m} \subset \mathbf{R}^d \times \{+1, -1\}^L$ be test set. $f_l(x)$ is the prediction function, and $rank_f(x, l) \in \{1, 2, \ldots, L\}$ is the ranking function.

1. Average Precision (AP): it is used to examine the average probability that the label whose position is ranked ahead of the predicted label for all samples still belongs to the sample label, defined as:

$$AP = \frac{1}{m}\sum_{i=1}^{m}\frac{1}{\left|R_i\right|} \times \sum_{l\in R_i}\frac{\left\{k\middle|rank_f\left(x_i,k\right) \le rank_f\left(x_i,l\right), k \in R_i\right\}}{rank_f\left(x_i,l\right)}.$$

2. Ranking Loss (RL): it is used to examine the average probability that the unrelated labels of all samples are ranked ahead of the related labels, defined as:

$$RL = \frac{1}{m}\sum_{i=1}^{m}\frac{1}{\left|R_i\right|\left|\bar{R}_i\right|} \times \left|\left\{(l,k)\middle|\text{rank}_f\left(x_i,l\right) \ge \text{rank}_f\left(x_i,k\right) \ , (l,k) \in R_i \times \bar{R}_i\right\}\right|.$$

3. Hamming Loss (HL): it is used to measure the misclassification of a sample on a single category label, defined as:

$$HL = \frac{1}{m}\sum_{i=1}^{m}\frac{1}{L}\sum_{l=1}^{L}\left[f_l\left(x_i\right) \ne Y_{il}\right].$$

4. Coverage (CV): it is a measure of the average number of lookups required for a sample to traverse all its relevant category labels, defined as:

$$CV = \frac{1}{m}\sum_{i=1}^{m}\max_{l\in R_i} rank_f\left(x_i,l\right) - 1.$$

5. One-Error (OE): it is a measure of the probability that the first label in the sample category label ranking is not part of the set of related labels, defined as:

$$OE = \frac{1}{m}\sum_{i=1}^{m}\left[\arg\max_{l\in L} f_l\left(x_i\right) \notin R_i\right],$$

where $R_i = \left\{l\middle|Y_{il} = +1\right\}$ denotes the set of labels related to sample x_i, and $\overline{R_i} = \left\{l\middle|Y_{il} = -1\right\}$ denotes the set of labels unrelated to sample x_i.

As for AP, the performance improves as the value increases, while for RL, HL, CV and OE, the performance improves as the value decreases.

Table 1 The description of multi-label datasets

Name	Instances	Attribute	Labels	Train	Test
Emotion	593	72	6	391	202
Birds	645	260	19	322	323
Yeast	2417	103	14	1499	918
Computer	5000	681	33	2000	3000
Health	5000	612	32	2000	3000

5.3 Experiment configurations

To validate the effectiveness of FIGR_LC, five classical multi-label feature selection algorithms are selected for comparison with the proposed algorithm. MLNB (Multi-label Naive Bayes Classification) [39] is a feature selection algorithm embedded in a multi-label Bayesian classifier. MDDM is a maximum dependency-based multi-label

feature selection algorithm, which is further divided into MDDMspc (MDDM Based on Subspace) [40] and MDDMproj (MDDM Based on Projection) [40] based on parameter selection. MEFS (Multi-label Embedded Feature Selection) [41] is an embedded feature selection method based on forecast risk. By evaluating the importance of each feature, the best feature subset is finally obtained. ARMLNRS [16] is a neighborhood rough set-based multi-label feature selection method. As for health and compute datasets with the feature number is more than 300, we take the first 300 features for the experiments. The multi-label classifier ML-*k*NN [42] is used to evaluate the performance of multi-label feature selection algorithms. *s* is the smoothing coefficient, which is set to 1. The number of nearest neighbours *k* is set to 10.

5.4 Experiment analysis and results

(1) *Comparative Performance* In order to demonstrate that FIGR_LC is effective, we compare the classification results based on different feature subsets induced by various feature selection algorithms, and analyse the classification performance of each feature selection algorithm with respect to the number of selected features. The experimental results are shown in Tables 2, 3, 4, 5 and 6.

As shown in the five tables, "↑" mark means that the performance improves as the value increases, while "↓" mark means that the performance improves as the value decreases. *Num* denotes the number of selected features, and it is equal to the number of features contained in the feature subset that could achieve the optimal classification result. In addition, bold indicates that the method obtains the best results for the relevant evaluation metric. "–" denotes that the method was unable to choose features because of its lengthy running time. Due to the failure of MEFS algorithm to get results on *health* and *computer* datasets, the classification results on these two datasets are replaced with "–". It is worth mentioning that the dimensionality reduction threshold of MDDM in the table is set to 90%, and MDDMspc, MDDMproj, MEFS and FIGR_LC could get the ranking of all the features, and then the top *Num* features that can achieve the

Table 2 Performance comparison on *emotion*

Algorithm	Num	AP (↑)	HL (↓)	CV (↓)	OE (↓)	RL (↓)
Raw data	72	0.781	0.214	1.920	0.332	0.173
MLNB	25	0.753	0.245	2.074	0.376	0.205
MDDMproj	3	0.669	0.314	2.649	0.445	0.316
MDDMspc	6	0.748	0.244	2.213	0.352	0.226
MEFS	69	0.798	0.219	1.891	0.297	0.166
ARMLNRS	30	0.790	0.226	1.960	**0.277**	0.174
FIGR_LC	45	**0.802**	**0.200**	**1.861**	0.282	**0.165**

Table 3 Performance comparison on *birds*

Algorithm	Num	AP (↑)	HL (↓)	CV(↓)	OE (↓)	RL (↓)
Raw data	260	0.695	0.054	3.399	0.390	0.125
MLNB	131	0.697	0.052	3.458	0.365	0.128
MDDMproj	13	0.626	0.068	3.564	0.523	0.141
MDDMspc	19	0.643	0.067	3.616	0.489	0.139
MEFS	128	0.709	0.053	3.511	0.344	0.129
ARMLNRS	104	0.719	0.053	3.421	0.346	0.123
FIGR_LC	111	**0.722**	**0.052**	**3.331**	**0.322**	**0.119**

Table 4 Performance comparison on *yeast*

Algorithm	Num	AP (↑)	HL (↓)	CV (↓)	OE (↓)	RL (↓)
Raw data	103	0.751	0.201	6.809	0.250	0.176
MLNB	28	0.736	0.208	6.693	0.256	0.187
MDDMproj	7	0.708	0.229	6.852	0.262	0.208
MDDMspc	10	0.712	0.229	6.879	0.254	0.206
MEFS	76	0.758	0.204	6.415	0.244	0.174
ARMLNRS	23	0.734	0.215	6.605	0.253	0.190
FIGR_LC	60	**0.758**	**0.201**	**6.330**	**0.237**	**0.171**

Table 5 Performance comparison on *health*

Algorithm	Num	AP (↑)	HL (↓)	CV (↓)	OE (↓)	RL (↓)
Raw data	612	0.681	0.046	3.305	0.421	0.061
MLNB	289	0.667	0.044	3.555	0.425	0.068
MDDMproj	15	0.634	0.049	3.834	0.491	0.074
MDDMspc	81	0.647	0.048	3.704	0.453	0.071
MEFS	–	–	–	–	–	–
ARMLNRS	83	0.685	0.043	3.358	0.401	0.063
FIGR_LC	57	**0.696**	**0.042**	**3.278**	**0.383**	**0.060**

Table 6 Performance comparison on *computer*

Algorithm	Num	AP (↑)	HL (↓)	CV (↓)	OE (↓)	RL (↓)
Raw data	681	0.633	0.041	4.416	0.437	0.092
MLNB	345	0.635	0.094	4.553	0.434	0.095
MDDMproj	18	0.598	0.044	4.880	0.481	0.105
MDDMspc	23	0.599	0.043	4.847	0.480	0.103
MEFS	–	–	–	–	–	–
ARMLNRS	132	0.633	0.040	4.419	0.443	0.091
FIGR_LC	189	**0.639**	**0.040**	**4.374**	**0.443**	**0.091**

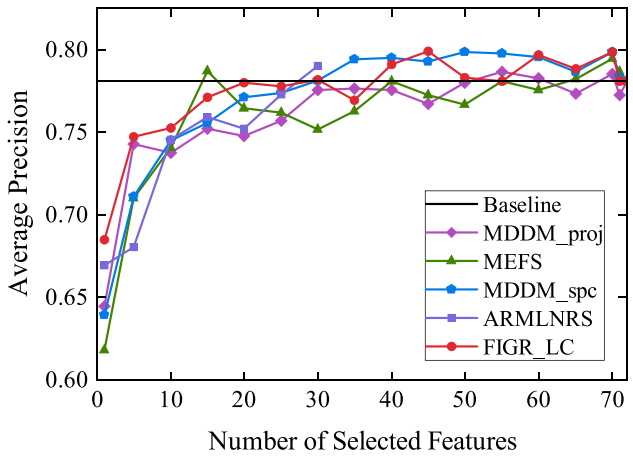

Fig. 1 Average precision of different algorithms on emotion

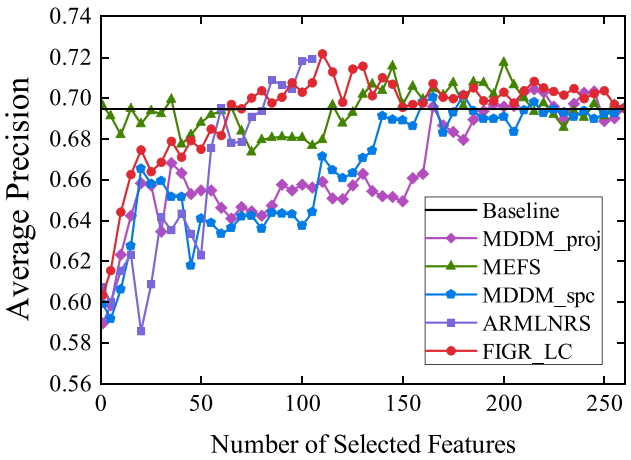

Fig. 2 Average precision of different algorithms on birds

best classification performance are selected to participate in the comparison.

From Tables 2, 3, 4, 5 and 6, it can be seen that FIGR_LC performs outstandingly compared to other algorithms, with optimal evaluation values on all datasets, except the *emotion* dataset where the OE metric is only slightly below the optimal perform. On the *emotion* dataset, ARMLNRS only achieves a winning rate of 3.33%, while FIGR_LC had a winning rate of 96.67%. According to the experimental results, the performance of FIGR_LC ranks first, followed by ARMLNRS and MEFS, and finally MLNB, MDDMproj and MDDMspc.

To analyze how varying numbers of selected features impact the algorithms' classification performance, we conducted further experiments, the outcomes of which are presented in Figs. 1, 2 and 3. The multi-label feature selection algorithms, namely MDDMspc, MDDMproj, MEFS, and FIGR_LC, generated a set of feature rankings, then inserted the features into the selected feature set sequentially according to the feature order. Subsequently, the resulting average classification accuracies were compared with ARMLNRS. The horizontal line in the figures represents the Average Precision (AP) of the initial data, providing a reference point for evaluation.

As can be seen from Figs. 1, 2 and 3, the algorithm performance fluctuates with the number of selected features. Nevertheless, the Average Precision (AP) does not follow a strictly increasing trend with the number of selected features.

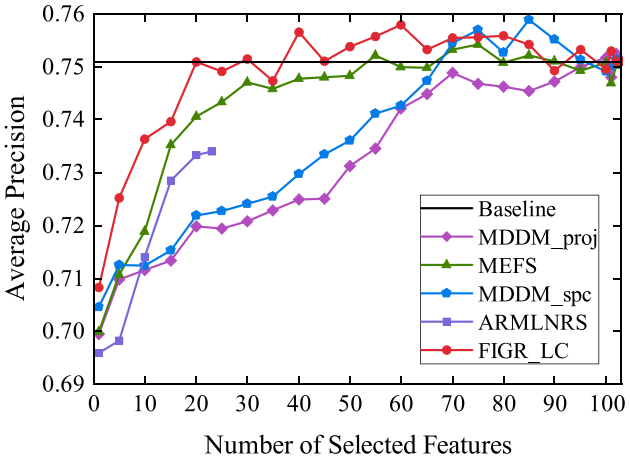

Fig. 3 Average precision of different algorithms on yeast

This suggests that not all features are equally effective; only specific ones influence the classification performance significantly. Notably, FIGR_LC displays markedly superior classification performance on these datasets compared to the other algorithms.

On the whole, MDDM exhibits two comparable mapping effects, with the average classification accuracy changing gradually as the selected feature number increases. When considering the maximum average Average Precision (AP), MDDMproj and MDDMspc require a selected feature number that is roughly similar to the original data's dimensionality, respectively. MEFS evaluates features using the classifier's scores, leading to an evident simplification effect, albeit with low efficiency. ARMLNRS selects fewer features, resulting in a noticeable simplification effect; however, the average classification accuracy remains relatively modest. In contrast, FIGR_LC not only achieves lower dimensionality but also attains an average classification accuracy that matches or even surpasses that of the original data, while also proving more efficiency than MEFS.

Analyzing the interplay between classification performance driven by feature subsets and the efficiency trend with increasing selected feature numbers provides a clearer demonstration of the proposed algorithm's effectiveness.

(2) *Statistical test* Statistical hypothesis tests were conducted to compare multiple multi-label feature selection algorithms across various datasets systematically, aiming to assess the efficiency of these methods. The Friedman test was employed to determine whether these comparative methods exhibited substantially differences. Given N datasets and κ comparative algorithms, $\mathfrak{r}_n^m$ denotes the ranking of the nth method on the mth dataset. $\mathfrak{R}_n = \frac{1}{N}\sum_{m=1}^{N}\mathfrak{r}_n^m$ indicates the nth method's average sort. This analysis was performed under the null hypothesis assuming that all methods are equal. The Friedman statistics are calculated as follows, where F_F follows the F-distribution with degrees of freedom $(\kappa-1)$ and $(\kappa-1)(N-1)$.

$$F_F = \frac{(N-1)\chi_F^2}{N(\kappa-1)-\chi_F^2}, \text{where}$$

$$\chi_F^2 = \frac{12N}{\kappa(\kappa+1)}\left(\sum_{n=1}^{\kappa}\mathfrak{R}_n^2 - \frac{\kappa(\kappa+1)^2}{4}\right).$$

In Table 7, F_F is presented, summarizing five evaluation metrics. The results reveal that at a significance level $\alpha = 0.1$, the null hypothesis is denied, indicating significant differences among the comparative approaches. A Nemenyi test was employed as a post-hoc test to assess whether FIGR_LC achieves competitive performance compared to the other algorithms. Comparing the difference between FIGR_LC and the average rating of a comparative algorithm is as follows:

$$CD = q_\alpha\sqrt{\frac{\kappa(\kappa+1)}{6N}}.$$

Given $q_\alpha = 2.589(\kappa = 6,\ \alpha = 0.1)$, $CD = 3.954(N = 3, \kappa = 6)$, the crucial difference CD is utilized to control the family-wise error rate.

Figure 4 illustrates the Critical Difference (CD) plots of several evaluation measures. In each subfigure, all methods' average ranks are aligned along the identical coordinate line. Methods with superior performance are located closer to the right edge of the axis, while those with lower rank are positioned closer to the left edge of the axis. A straight line connects any two methods if their average ranks are less than the CD. Methods whose average ranks exceeds the CD are considered to have substantial distinctions.

In summary, it can be seen from Fig. 4 that: (1) FIGR_LC outperforms MDDMspc, MDDMproj, MLNB, MEFS and ARMLNRS across all evaluation metrics. (2) FIGR_LC demonstrates statistically superior over MDDMproj across all evaluation measures.

Table 7 Summary of the Friedman statistics F_F $(k = 6,\ N = 3)$ and the critical value

Evaluation measure	F_F	Critical value $(\alpha = 0.10)$
AP	29.5	2.522
RL	17.6875	
HL	7.403	
CV	17.6875	
OE	17.6875	

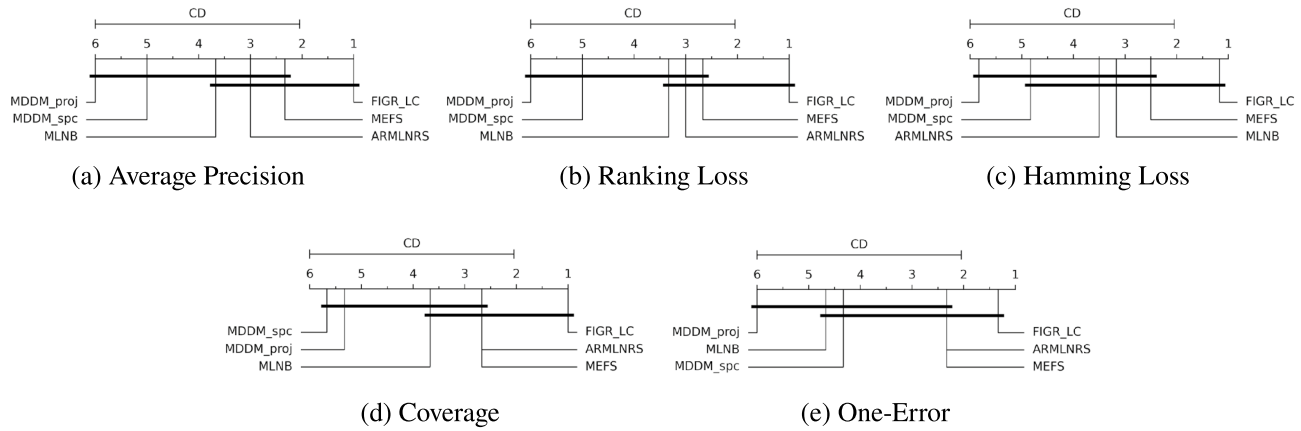

Fig. 4 Comparison of FIGR_LC against other comparing algorithms with the Nemenyi test

6 Conclusions

To address the inherent ambiguity and uncertainty in multi-label data, this paper introduces a novel multi-label feature selection algorithm utilizing fuzzy information gain ratio. This approach simultaneously considers the correlation between features and individual labels, the correlation between features and the set of labels, as well as the correlation between labels when evaluating feature importance. Comparative analysis with other existing multi-label feature selection algorithms demonstrates the superiority of the proposed algorithm.

In traditional multi-label learning, it is commonly assumed that each training sample is accurately labeled with all relevant labels. However, this assumption rarely holds true in reality due to the exorbitant cost associated with precise labeling for each sample. Instead, just roughly assigning a set of candidate labels to each sample would significantly reduce labeling efforts, leading to the emergence of Partial Multi-label Learning (PML). Although PML is a prominent research area within multi-label learning, the feature selection of partial multi-label learning remains underexplored. In our future work, we plan to explore several effective partial multi-label feature selection algorithms based on fuzzy information entropy.

Acknowledgements The authors would like to thank the Editors for their kindly help and the anonymous referees for their valuable comments and helpful suggestions. The work is partially supported by the National Natural Science Foundation of China (Serial No. 62163016, 62066014), the Natural Science Foundation of Jiangxi Provincial (Serial No. 20212ACB202001, 20232BAB202004), the open project of State Key Laboratory of Performance Monitoring and Protecting of Rail Transit Infrastructure, East China Jiaotong University (Grant No. HJGZ2023203), and the Jiangxi Double Thousand Plan.

Data availability The data that support the findings of this study are openly available in *Mulan* at https://mulan.sourceforge.net/.

References

1. Mitchell TM, Mitchell TM (1997) Machine learning. McGraw-Hill, New York
2. Zhang M-L, Zhou Z-H (2013) A review on multi-label learning algorithms. IEEE Trans Knowl Data Eng 26(8):1819–1837
3. Tsoumakas G, Katakis I (2007) Multi-label classification: an overview. Int J Data Wareh Min 3(3):1–13
4. Li L, Wang M, Zhang L, Wang H (2014) Learning semantic similarity for multi-label text categorization. In: Workshop on Chinese lexical semantics. Macao, China, pp 260–269
5. Jiang J-Y, Tsai S-C, Lee S-J (2012) Fsknn: multi-label text categorization based on fuzzy similarity and k nearest neighbors. Expert Syst Appl 39(3):2813–2821
6. Wang C, Yan S, Zhang L, Zhang H-J (2009) Multi-label sparse coding for automatic image annotation. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). FL, USA, Miami, pp 1643–1650
7. Wu B, Lyu S, Hu B-G, Ji Q (2015) Multi-label learning with missing labels for image annotation and facial action unit recognition. Pattern Recognit 48(7):2279–2289
8. Yu Y, Pedrycz W, Miao D (2013) Neighborhood rough sets based multi-label classification for automatic image annotation. Int J Approx Reason 54(9):1373–1387
9. Zhao F, Huang Y, Wang L, Tan T (2015) Deep semantic ranking based hashing for multi-label image retrieval. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). MA, USA, Boston, pp 1556–1564
10. Yu K, Yu S, Tresp V (2005) Multi-label informed latent semantic indexing. In: Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '05). Salvador, Brazil, pp 258–265
11. Zhu C, Liu Y, Miao D, Dong Y, Pedrycz W (2023) within-cross-consensus-view representation-based multi-view multi-label learning with incomplete data. Neurocomputing 557:126729
12. Zhu C, Miao D, Wang Z, Zhou R, Wei L, Zhang X (2020) global and local multi-view multi-label learning. Neurocomputing 371:67–77
13. Lin Y, Hu Q, Liu J, Chen J, Duan J (2016) Multi-label feature selection based on neighborhood mutual information. Appl Soft Comput 38:244–256
14. Li L, Liu H, Ma Z, Mo Y, Duan Z, Zhou J, Zhao J (2014) Multi-label feature selection via information gain. In: International

conference on advanced data mining and applications (ADMA). Guilin, China, pp 345–355
15. Gao C, Zhou J, Xing J, Yue X (2022) parameterized maximum-entropy-based three-way approximate attribute reduction. Int J Approx Reason 151:85–100
16. Duan J, Hu Q, Zhang L, Qian Y, Li D (2015) Feature selection for multi-label classification based on neighborhood rough sets. J Comput Res Dev 52(1):56–65
17. Li J, Yang X, Wang P, Chen X (2018) Stable attribute reduction approach for fuzzy rough set. J Nanjing Univ Sci Technol (Nanjing Li Gong Daxue Xuebao) 42(1):68–75
18. Reyes O, Morell C, Ventura S (2015) Scalable extensions of the relieff algorithm for weighting and selecting features on the multi-label learning context. Neurocomputing 161:168–182
19. Dai J, Hu Q, Hu H, Huang D (2017) Neighbor inconsistent pair selection for attribute reduction by rough set approach. IEEE Trans Fuzzy Syst 26(2):937–950
20. Pawlak Z (1982) Rough set. Int J Comput Inf Sci 11(5):341–356
21. Zhang C, Li D, Liang J (2018) Hesitant fuzzy linguistic rough set over two universes model and its applications. Int J Mach Learn Cybern 9(4):577–588
22. Qian Y, Liang J, Pedrycz W, Dang C (2010) Positive approximation: an accelerator for attribute reduction in rough set theory. Artif Intell 174(9–10):597–618
23. Yao Y, Zhang X (2017) Class-specific attribute reducts in rough set theory. Inf Sci 418:601–618
24. Liang J, Wang F, Dang C, Qian Y (2014) A group incremental approach to feature selection applying rough set technique. IEEE Trans Knowl Data Eng 26(2):294–308
25. Qian Y, Wang Q, Cheng H, Liang J, Dang C (2015) Fuzzy-rough feature selection accelerator. Fuzzy Sets Syst 258:61–78
26. Wang C, Wang Y, Shao M, Qian Y, Chen D (2019) Fuzzy rough attribute reduction for categorical data. IEEE Trans Fuzzy Syst 28(5):818–830
27. Zhao H, Wang P, Hu Q, Zhu P (2019) Fuzzy rough set based feature selection for large-scale hierarchical classification. IEEE Trans Fuzzy Syst 27(10):1891–1903
28. Ni P, Zhao S, Wang X, Chen H, Li C, Tsang EC (2020) Incremental feature selection based on fuzzy rough sets. Inf Sci 536:185–204
29. Hu Q, Yu D, Xie Z, Liu J (2006) Fuzzy probabilistic approximation spaces and their information measures. IEEE Trans Fuzzy Syst 14(2):191–201
30. Zhang X, Mei C, Chen D, Yang Y, Li J (2019) Active incremental feature selection using a fuzzy-rough-set-based information entropy. IEEE Trans Fuzzy Syst 28(5):901–915
31. Jensen R, Shen Q (2009) New approaches to fuzzy-rough feature selection. IEEE Trans Fuzzy Syst 4(17):824–838
32. Jensen R, Shen Q (2007) Fuzzy-rough sets assisted attribute selection. IEEE Trans Fuzzy Syst 15(1):73–89
33. Zhang L, Hu Q, Duan J, Wang X (2014) Multi-label feature selection with fuzzy rough sets. In: International conference on Rough Sets and Knowledge Technology (RSKT). Shanghai, China, pp 121–128
34. Lin Y, Hu Q, Liu J, Li J, Wu X (2017) Streaming feature selection for multilabel learning based on fuzzy mutual information. IEEE Trans Fuzzy Syst 25(6):1491–1507
35. Lin Y, Li Y, Wang C, Chen J (2018) Attribute reduction for multi-label learning with fuzzy rough set. Knowl Based Syst 152:51–61
36. Li Y, Lin Y, Liu J, Weng W, Shi Z, Wu S (2018) Feature selection for multi-label learning based on kernelized fuzzy rough sets. Neurocomputing 318:271–286
37. Dubois D, Prade H (1990) Rough fuzzy sets and fuzzy rough sets. Int J Gen Syst 17(2–3):191–209
38. Dai J, Chen J, Liu Y, Hu H (2020) Novel multi-label feature selection via label symmetric uncertainty correlation learning and feature redundancy evaluation. Knowl Based Syst 207:106342
39. Zhang M-L, Peña JM, Robles V (2009) Feature selection for multi-label naive bayes classification. Inf Sci 179(19):3218–3229
40. Zhang Y, Zhou Z-H (2008) Multi-label dimensionality reduction via dependence maximization. ACM Trans Knowl Discovery Data 4(3):1–21 (Article No. 14)
41. Ge L, Li G, You M (2009) Embedded feature selection for multi-label learning. J Nanjing Univ (Nat Sci) 45(5):671–676
42. Zhang M-L, Zhou Z-H (2007) Ml-knn: a lazy learning approach to multi-label learning. Pattern Recognit 40(7):2038–2048